
Institute for Defense Analyses

Formulation of Default Correlation Values for Cost Risk Analysis

Bruce Harmon

Presented at SSCAG/SCAF/EACEWG Joint Meeting
May 11–12, 2010 Berlin, Germany





Importance of Correlation in Cost/Risk Analysis

- Most program cost estimates are created by summing over multiple work breakdown structure (WBS) elements
- Statistical properties of a sum are dependent on correlations between the summed elements.
- The higher the correlations, the greater the dispersion in estimates of the sum.
- As WBS element correlations are generally positive ($\rho > 0$), ignoring correlation (assuming $\rho = 0$) results in an underestimate of dispersion.



What if the Analyst has no Knowledge of Correlation Values?

- Analyses by Book¹ suggest a default value of $\rho=.2$
- The Book heuristic is employed and cited by the space cost estimating community
- IDA used Book's analysis as a jumping-off point for formulating alternative default correlation values

—

¹Stephen A. Book, “Why Correlation Matters in Cost Estimating”, 32nd Annual DoD Cost Analysis Symposium, Williamsburg, VA, February 1999

- Start with formula for variance of a sum:

$$Var(C) = \sum_{i=1}^n Var(C_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \rho_{ij} \sqrt{Var(C_i)Var(C_j)} .$$

- Make simplifying assumptions for sensitivity analyses
 - All element variances and correlations are equal and non-negative
 - If $\rho = 0$, $Var(C) = nVar(C_i) = n\sigma_i^2$
 - Where $\rho > 0$, $Var(C) = n\sigma_i^2 + \rho(n^2 - n)\sigma_i^2$



Book "Knee in the Curve" Relationship

- Equation describing inaccuracy in σ when the analyst specifies no correlation ($\rho=0$) but the true correlation is positive ($\rho>0$)

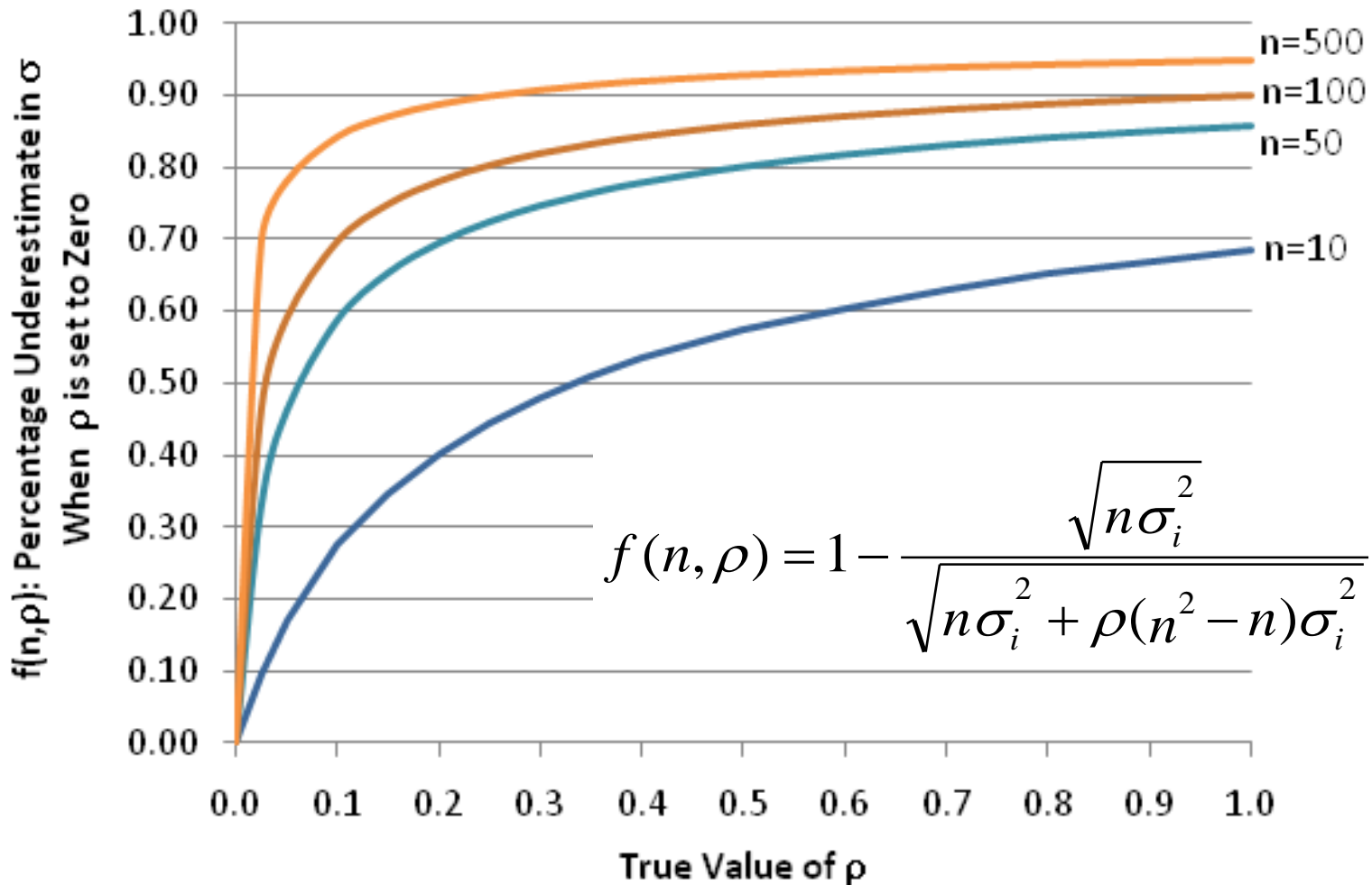
$$f(n, \rho) = 1 - \frac{\sqrt{n\sigma_i^2}}{\sqrt{n\sigma_i^2 + \rho(n^2 - n)\sigma_i^2}}$$

σ given no correlation

Incremental variance due to correlation

- Interpreted as percentage underestimate in σ when correlation is ignored but is positive

Knee in the Curve at $\rho=.2$



Further Analysis by Book

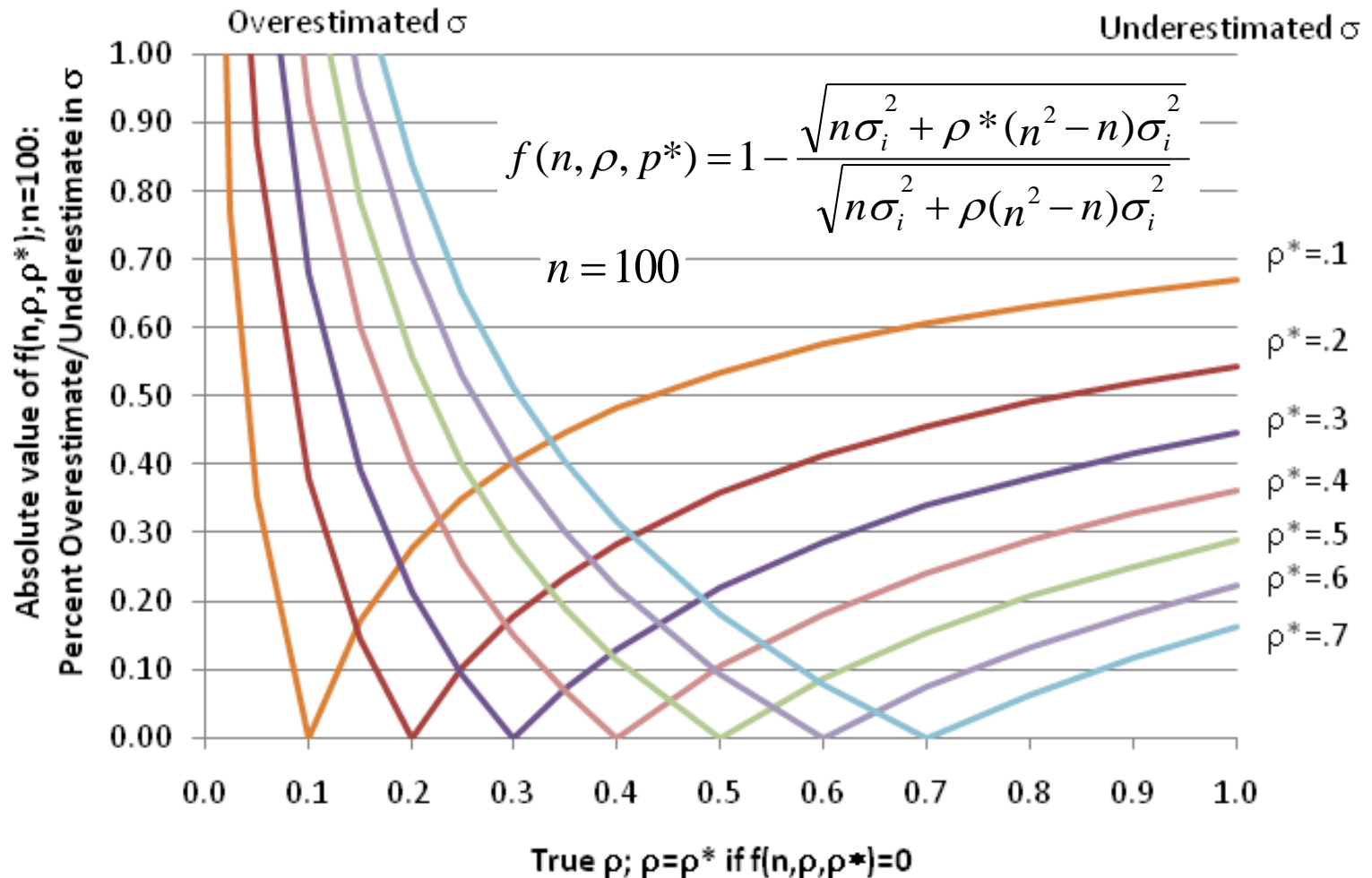
- Book presents a modified function where the analyst's choice of ρ (ρ^*) is varied around $\rho^* = .2$

$$f(n, \rho, \rho^*) = 1 - \frac{\sqrt{n\sigma_i^2 + \rho^*(n^2 - n)\sigma_i^2}}{\sqrt{n\sigma_i^2 + \rho(n^2 - n)\sigma_i^2}}$$

- In this case percentage errors can be positive or negative
 - Graphical representations of $f(n, \rho, \rho^*)$ are expressed as absolute values
- Visual inspection indicates balanced over and under estimates at $\rho^* = .2$



Sensitivity of Percent Error in σ When Choice of ρ^* is Varied





Alternative Formulation

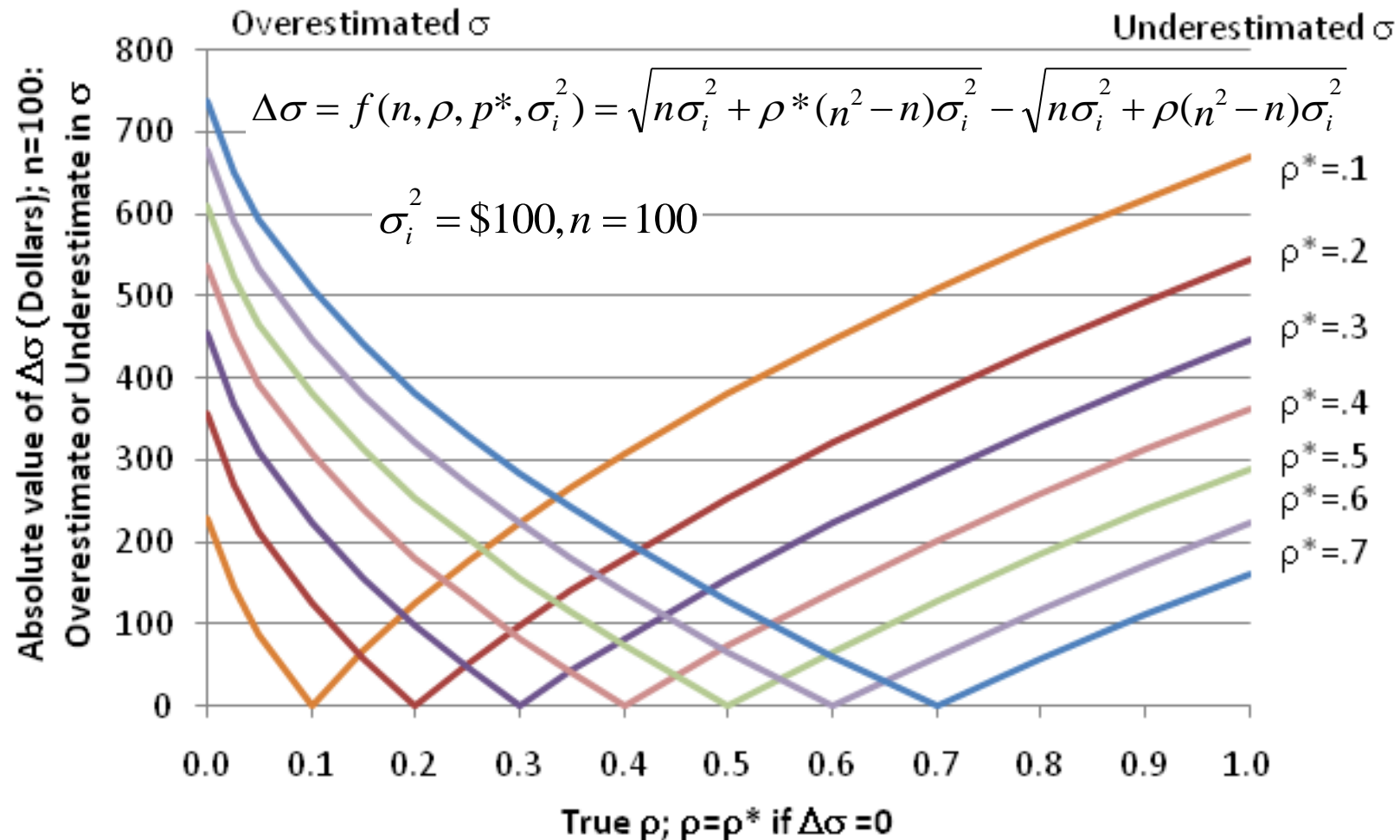
- Book recommends $\rho^* = .2$ based on perceived balance of percentage errors in σ
- As σ is in the dimension of the cost estimate, formulate function in terms of raw error

$$\Delta\sigma = f(n, \rho, \rho^*, \sigma_i^2) = \sqrt{n\sigma_i^2 + \rho^*(n^2 - n)\sigma_i^2} - \sqrt{n\sigma_i^2 + \rho(n^2 - n)\sigma_i^2}$$

- Find ρ^* where the expected value of $\Delta\sigma$ is zero:
 $E(\Delta\sigma) = 0$
 - Explicitly balance over and underestimates



Sensitivity of $|\Delta\sigma|$ When ρ^* is Varied



- **Implied distribution of ρ is uniform**

- The analyst has no priors for ρ (other than non-negativity)

- $$f(x) = \frac{1}{b-a}, a=0, b=1.$$

- **Given this, we can derive $E(\Delta\sigma)$:**

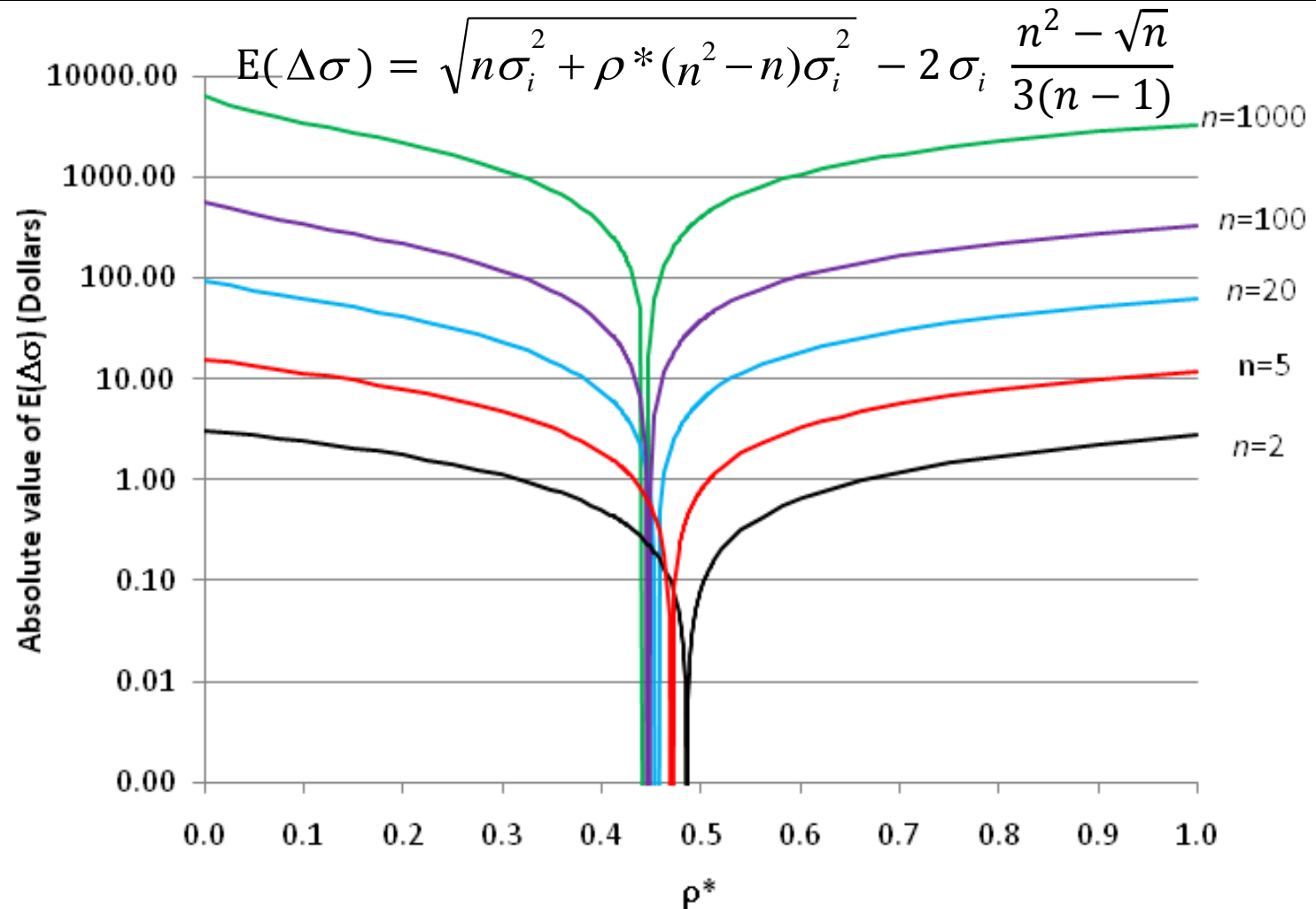
As $\sqrt{n\sigma_i^2 + \rho^*(n^2 - n)\sigma_i^2}$ is not affected by ρ , we only need to derive the expected value of $\sqrt{n\sigma_i^2 + \rho(n^2 - n)\sigma_i^2}$

$$E(g(x)) = \int_0^1 \sqrt{n\sigma_i^2 + x(n^2 - n)\sigma_i^2} dx = 2\sigma_i \frac{n^2 - \sqrt{n}}{3(n-1)};$$

$$E(\Delta\sigma) = \sqrt{n\sigma_i^2 + \rho^*(n^2 - n)\sigma_i^2} - 2\sigma_i \frac{n^2 - \sqrt{n}}{3(n-1)}$$



Sensitivity of $|E(\Delta\sigma)|$ to ρ^*





Find $E(\Delta\sigma)=0$ for a given n

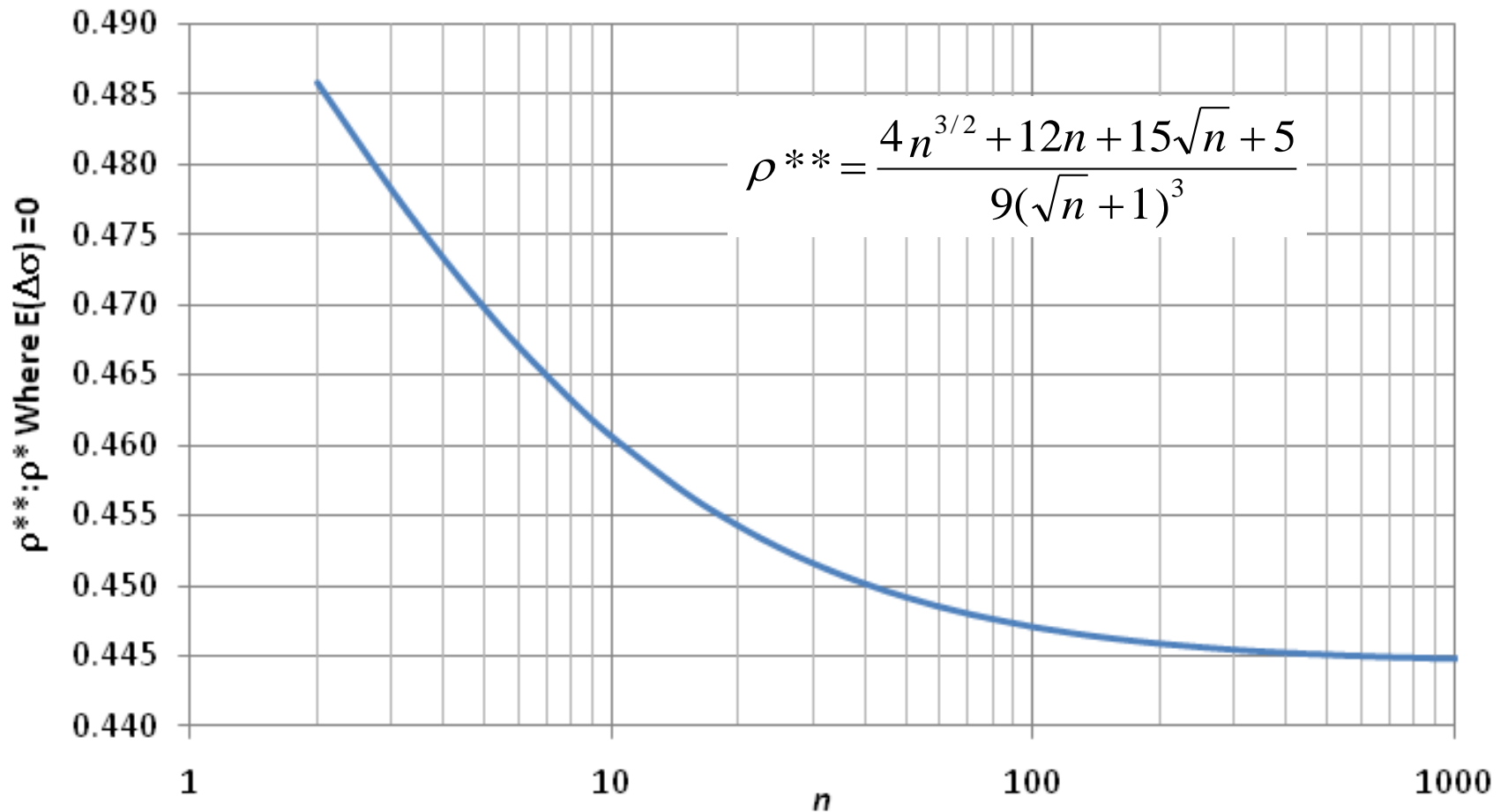
- Optimum ρ^* (ρ^{**}) is where $E(\Delta\sigma)=0$
- Solve for ρ^*

$$E(\Delta\sigma) = \sqrt{n\sigma_i^2 + \rho^*(n^2 - n)\sigma_i^2} - 2\sigma_i \frac{n^2 - \sqrt{n}}{3(n-1)} = 0$$

$$\rho^{**} = \frac{4n^{3/2} + 12n + 15\sqrt{n} + 5}{9(\sqrt{n} + 1)^3}$$

- Note that σ is not included in this expression

Sensitivity of ρ^{**} to n





Conclusions

- If the analyst has no prior knowledge of correlations (but thinks they are positive), a default value of around .45 is appropriate
- The methodology can be applied to other prior beliefs regarding correlation bounds
 - e.g the correlations fall between $-.2$ and 1

–

–